Application of artificial intelligence technology in unconventional natural gas production forecasting

Qichao Gao[a], Lulu Liao[*ab], Shunhui Yang[a]
[a]SINOPEC Research Institute of Petroleum Engineering Co. Ltd., Beijing, China; [b]College of Petroleum Engineering, China University of Petroleum, Beijing, China

ABSTRACT

Improving the speed and accuracy of unconventional natural gas production forecasts is the key to scientific and efficient development of unconventional resources. The existing prediction methods based on the transmission mechanism make assumptions and simplifications on the model, and it is difficult to comprehensively and accurately evaluate the main control factors of production capacity, resulting in large production prediction errors. This paper proposes a productivity prediction method for unconventional natural gas wells based on artificial intelligence (AI) and data mining technologies. We use the Pearson correlation coefficient and grey relational analysis to screen out the main control factors, and select the best yield intelligent forecasting model by training and comparing a variety of commonly used machine learning methods. This paper takes the Alberta tight gas field in Canada as an example to illustrate the effectiveness and practicability of this method.

Keywords: Artificial intelligence, data mining, unconventional natural gas, production foresting

1. INTRODUCTION

Traditional production modeling and forecasting methods are mostly based on transport mechanisms of shale gas, such as through simulation models and theoretical analysis[1]. These methods are based on certain assumptions and simplifications, and based on seepage theory, establish differential equations to solve the productivity formula, and these methods have high computational efficiency[2]. For reservoirs with complex mechanisms and strong heterogeneity, the established differential equations cannot be analytically solved, and numerical methods can be used to solve the productivity. For reservoirs with complex seepage mechanisms (such as tight gas, shale gas, etc.), the traditional productivity prediction methods will deviate from the production prediction results due to incomplete mechanism consideration and ideal assumptions of the model[3]. In addition, when numerical simulations are used to calculate production, the models rely heavily on accurate geological models, resulting in prolonged modeling cycles and reduced efficiency.

With the rise of the third wave of AI technology revolution, relevant data mining and intelligent techniques provide new avenues for production forecasting of unconventional natural gas[4]. Compared with the traditional mechanism-driven method, the data-driven artificial intelligence method can establish a proxy model related to the production, geological reservoir and fracturing operation parameters through deep mining of oilfield data[5]. The current large-scale drilling and fracturing data of unconventional reservoirs provides a possible scenario for the large-scale application of artificial intelligence methods[6].

This study first introduces the workflow of the development of intelligence prediction model, and then explains the data mining and artificial intelligence methods, technologies and principles behind it, and finally takes the northern Alberta tight gas field in Canada as an example to illustrate the artificial intelligence workflow in data processing, feature selection, parameter optimization, and model optimization in unconventional natural gas production forecasting. This study provides a basis for the popularization and application of artificial intelligence methods in unconventional natural gas production forecasting.

[*] lulu.liao2000@hotmail.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066R
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662821

2. WORKFLOW OF INTELLIGENT FORECASTING MODEL DEVELOPMENTS

The process of using artificial intelligence technology to build an intelligent production forecast model includes 5 steps, as shown below.

(1) Raw data collection and preprocessing. The original data set includes characteristic parameters and production labels. Data preprocessing operations include outlier data removal, data padding, data normalization, and normalization[7].

(2) Feature optimization. We use the Person correlation coefficient to find the correlation of variables, and then using the gray correlation to analyze the weight of the main controlling factors affecting the production capacity.

(3) Data set division. The original dataset is usually divided using the set-out method. According to a set ratio, the original dataset is randomly divided into training set and test set.

(4) Artificial intelligence modeling. Selecting the learner and then training the learner with the training dataset generated in the previous step, this study tested a variety of learners for modeling, including decision forests, support vector machines, and random forest and so on[8].

(5) Model evaluation. This study input the test set data into the trained model, calculate and analyze the error and accuracy of the predicted production. The usual evaluation indicators include the RMSE (the root mean square error), the MAE (mean absolute error) and the R^2 (coefficient of determination). In this study, RMSE was used to evaluate the production forecasting model[9].

3. THE PRINCIPLES AND METHODS OF ARTIFICIAL INTELLIGENCE MODELING

The methodology and principles of AI agent model development are discussed in detail for workflow.

3.1 Data set acquisition and preprocessing

Engineering geological parameters and fracturing construction data were collected from the field and used to construct the structured dataset. The original dataset cannot be directly used to train the model, and a series of preprocessing operations are required. In the original dataset, when the missing data in the feature variable is greater than the safety threshold of 5%, it will not be considered in the modeling. The original data needs to be normalized as different units may affect the results[10]. The Z-Score method is an ideal solution and is often used in data normalization. It represents how many standard deviations the measurement data deviates from the mean of the data population, and its expression is:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where z is the standardized dimensionless data; σ is the sample standard deviation; x is the original value of a sample; μ is the mean value of the sample. Assuming the dataset D contains n learning samples, each sample has m feature parameters and a feature label, namely**:**

$$D = \{x_{1i}, x_{1i}, ..., x_{mi}, y_i\} \quad \left(i = 1, 2, ..., n; n > m\right) \tag{2}$$

3.2 Feature optimization

There are many methods for feature optimization. Commonly used methods include multiple regression method, gray correlation method, Lasso method, covariance and so on[11]. The Pearson correlation coefficient characterizes the similarity between variables, and its output ranges from -1 to +1, among them, 0 means that the vectors are independent of each other, and the closer the absolute value is to 1, the stronger the correlation[12]. The formula for calculating the Pearson correlation coefficient is as follows:

$$\rho\left(x_1, x_2\right) = \frac{cov\left(x_1, x_2\right)}{\sigma_{x_1}\sigma_{x2}} \tag{3}$$

where *cov* is the covariance and *σ* is the standard deviation. Compared with covariance, Pearson's correlation coefficient removes the effect of variable dimension.

This study uses the Grey Relation Analysis (GRA) method to study the influence of the selected features, and the grey relational analysis of the main controlling factors helps to enhance the interpretability of the model. The GRA method is a multivariate statistical analysis research tool. Simply put, in a gray system, we want to know how strongly a variable is affected by other variables or factors[13]. Grey relational degree analysis has low requirements on data and calculation, and its calculation formula is as follows:

$$\xi_i\left(k\right)=\frac{\min\limits_{i}\min\limits_{k}\left|x_0\left(k\right)-x_i\left(k\right)\right|+\rho\max\limits_{i}\max\limits_{k}\left|x_0\left(k\right)-x_i\left(k\right)\right|}{\left|x_0\left(k\right)-x_i\left(k\right)\right|+\rho\max\limits_{i}\max\limits_{k}\left|x_0\left(k\right)-x_i\left(k\right)\right|}\quad\left(k=1,2,...,m\right)\tag{4}$$

where *i* is the serial number of a sample; *k* represents the serial number of a feature; *ρ* is the resolution coefficient, ranging from 0 to 1. The smaller the *ρ*, the more obvious the difference between samples, and usually *ρ* is taken as 0.5. $x_0(k)$ is selected by user and regarded as the reference sequence. According to the calculated correlation coefficient, the influence of each factor on the production can be analyzed. The correlation order can be further calculated based on the correlation coefficient, and the formula is:

$$r_{0i}=\frac{1}{m}\sum_{k=1}^{m}\varsigma_i\left(k\right)\tag{5}$$

The degree of influence of each factor can be obtained by sorting the correlation degree.

3.3 Data set division

The original dataset is usually divided using the set-out method and setting a ratio. When the amount of data is relatively small, a ratio of 7:3 to 9:1 can usually be used to divide training data and test data. Commonly used partitioning methods for datasets include hand-out method and cross-validation method. The hand-out method is to use a portion of the total samples as test set. The proportion is usually 10%-30%, and generally the capacity of the test set is at least more than 30.

To ensure that the model's predictions are reliable, this study performed 300 samplings, divided the training and test sets, and trained and evaluated the selected machine learning models accordingly[14].

3.4 Intelligent algorithm model selection

Commonly used intelligent algorithms are compared in this study, including gradient boosting, random forests, support vectors, and neural networks. The training set data is used to train the intelligent model, and the performance of the model is evaluated by using the test set[15-17].

For regression problems, commonly model prediction evaluation indicators include mean square error, root mean square error, mean absolute error, coefficient of determination, etc. In this study, RMSE (Root Mean Square Error) is used as a key indicator to compare the performance of different models, and its expression is as follows:

$$RMSE(X)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}\left(x_i\right)-y_i)^2}\tag{6}$$

Accuracy is also used as a key indicator to compare the performance of different models, and its expression is as follows:

$$acc=\left|\frac{y_{pred}-y_{real}}{y_{real}}\right|\times100\%\tag{7}$$

4. APPLICATION CASE

4.1 Data collection and preprocessing

This paper takes the W tight gas reservoir in Alberta, Canada as a case study. The depth of the W reservoir is 2000 meters. The research target is located in the RT21 block, with an area of 2500 km^2 and an average reservoir thickness of 200 meters. The samples of 1091 wells in the study area were de-noised, cleaned and screened, and there were 1071 available data sample wells, and each sample contained 12 dimension features and 1 set of production label. After missing vale pre-processing, the feature dimension of the model is changed from 12 to 10.

4.2 Feature optimization

If the feature parameters of the model input are strongly correlated, it will not only increase the model training time, but also affect the interpretability of the model. Figure 1 shows the Pearson correlation coefficient matrix of the characteristic variables. Total Proppant Pumped dose has a strong linear positive correlation with Proppant Pumped Per Stimulated Length (Pearson's coefficient is 0.92). Total Fluid also has a strong linear positive correlation with Fluid Per Stimulated Length (Pearson's coefficient is 0.86). After removing the linear correlation variables Proppant Pumped Per Stimulated Length and Fluid Per Stimulated Length, the dimension of the feature variable is reduced from 10 dimensions to 8 dimensions.

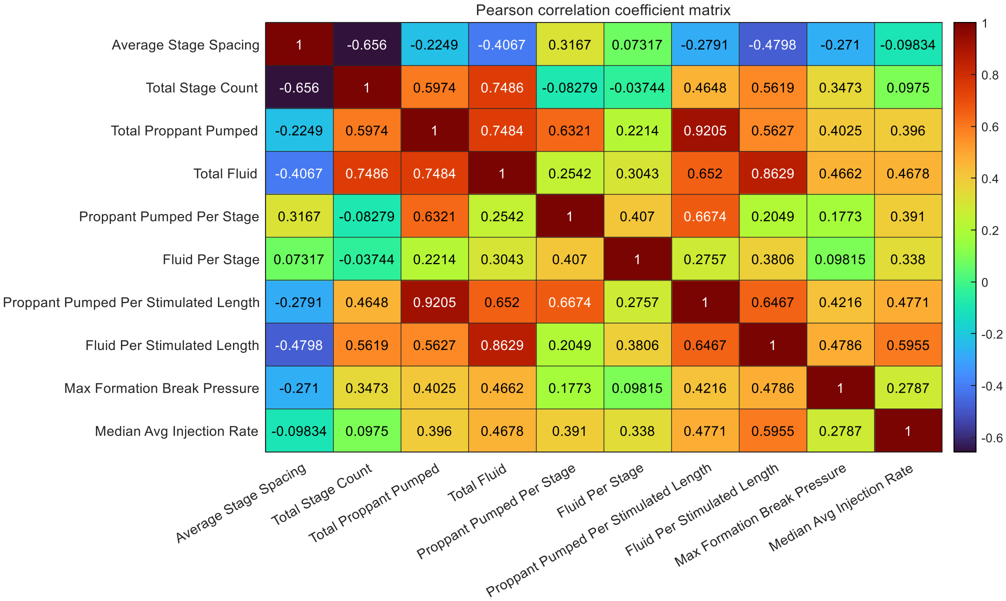

Figure 1. Variable correlation analysis matrix.

The grey correlation method can be used to calculate the correlation between the 8 control factors and the production. As shown in figure 2, it can be seen that Fluid Per Stage has the greatest impact on the Production and Total Proppant Pumped has the least impact on the production.

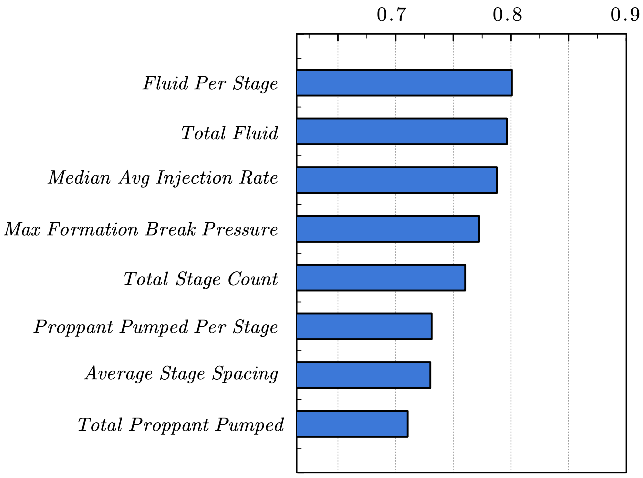

Figure 2. Tornado chart of controlling factors by grey correlation method.

4.3 AI model optimization and analysis

In order to select the optimal model, this study evaluates gradient boosting, decision tree, random forest, support vector machine and neural network respectively. The evaluation results are shown in table 1 below.

Table 1. Intelligent model predictive evaluation.

Models	RMSE	Test set accuracy	Running time	Train: Test	Training times
Gradient boosting	21.3	82.08 %	2.78 s	8:2	300
Decision tree	30.24	77.23 %	1.35 s	8:2	300
Random forest	10.9	85.25 %	12.63 s	8:2	300
Support vector machine	25.68	83.99 %	88.58	8:2	300
Neural networks	28.32	81.2%	72.68	8:2	300

From the table 1, we can acknowledge that the performance of the random forest model is the best, the test set accuracy can reach more than 85%, and the model training time is about 12.63 s, which is relatively short. Therefore, the random forest is chosen as the intelligence model in this research area.

5. CONCLUSIONS

This study proposes a productivity prediction method and workflow for unconventional natural gas wells based on data mining and artificial intelligence technology. Through automated data pre-processing, sample sampling, feature analysis, and model screening, the rapid and efficient prediction of unconventional natural gas productivity is achieved. The paper takes the Canadian gas field as an example to demonstrate the specific application and effect of artificial intelligence technology in unconventional natural gas production forecasting. The following conclusions can be drawn from the intelligent model study for unconventional natural gas:

(1) The artificial intelligence prediction method is an effective supplement to the existing mechanism-driven prediction method. It has comprehensive analysis capabilities and can quickly and efficiently predict production.

(2) Grey relational analysis helps to enhance the interpretability of the model, and it is clear that the main controlling factor affecting the study area are Fluid Per Stage, Total Fluid etc.

(3) The Pearson correlation coefficient matrix can analyze the correlation of sample features, effectively reduce the dimension of data set features, and reasonably simplify the model.

(4) Using the RMSE, test accuracy and training time indicators, the optimal intelligent model in the study area can be effectively selected as the random forest model.

ACKNOWLEDGMENTS

This study is supported by the National Key R&D Program (approval number: 2019YFA0708300). The authors thank colleagues at SINOPEC Research Institute of Petroleum Engineering and scholars from China University of Petroleum (Beijing) for their support. The author also sincerely thanks the editors and reviewers for editing and reviewing this manuscript.

REFERENCES

[1] Prieto, M., Aristizabal, J. A., Pradilla, D. and Gómez, J. M., "Simultaneous numerical simulation of the hydraulic fractures geometry in multi-stage fracturing for horizontal shale gas wells," Journal of Natural Gas Science and Engineering 102(52), 104567 (2022).

[2] Gao, Q., Dong, P. and Liu, C., "Study on the influence of shale storage space types on shale gas transport," ACS omega 6(20), 12931-12951 (2021).
[3] Gao, Q., Dong, P. and Liu, C., "Modeling and simulation of shale fracture attitude," ACS omega 6(11), 7312-7333 (2021).
[4] Han, J., Pei, J. and Tong, H., "Getting to know your data," Data Mining, 29-38 (2022).
[5] Gao, X., Dong, P., Cui, J. and Gao, Q., "Prediction model for the viscosity of heavy oil diluted with light oil using machine learning techniques," Energies 15(6), 2297 (2022).
[6] Liao, L., Zeng, Y., Liang, Y. and Zhang, H., "Data mining: A novel strategy for production forecast in tight hydrocarbon resource in Canada by random forest analysis," International Petroleum Technology Conference, 1-9 (2020).
[7] Mishra, P., Biancolillo, A., Roger, J. M., Marini, F. and Rutledge, D. N., "New data preprocessing trends based on ensemble of multiple preprocessing techniques TrAC," Trends in Analytical Chemistry 132(56), 116045 (2020).
[8] Zhu, H., "Big data and artificial intelligence modeling for drug discovery," Annual Review of Pharmacology and Toxicology 60(15), 573 (2020).
[9] Wang, W. and Lu, Y., "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," IOP Conference Series: Materials Science and Engineering, 12049 (2018).
[10] Curtis, A. E., Smith, T. A., Ziganshin, B. A. and Elefteriades, J. A., "The mystery of the Z-score," Aorta 4(04), 124-130 (2016).
[11] Gauraha, N., "Introduction to the LASSO," Resonance 23(4), 439-464 (2018).
[12] Edelmann, D., Móri, T. F. and Székely, G. J., "On relationships between the Pearson and the distance correlation coefficients," Statistics & Probability Letters 169(32), 108960 (2021).
[13] Si, A., Das, S. and Kar, S., "Picture fuzzy set-based decision-making approach using Dempster–Shafer theory of evidence and grey relation analysis and its application in COVID-19 medicine selection," Soft Computing, 1-15 (2021).
[14] Bell, J., "What is machine learning?," Machine Learning and the City: Applications in Architecture and Urban Design 28(16), 207-216 (2022).
[15] Khan, M. A., Shah, M. I., Javed, M. F., Khan, M. I., Rasheed, S., El-Shorbagy, M. A., El-Zahar, E. R. and Malik, M. Y., "Application of random forest for modelling of surface water salinity," Ain Shams Engineering Journal 13(4), 101635 (2022).
[16] Wright, L. G., Onodera, T., Stein, M. M., Wang, T., Schachter, D. T., Hu, Z. and McMahon, P. L., "Deep physical neural networks trained with backpropagation," Nature 601(7894), 549-555 (2022).
[17] Essam, Y., Huang, Y. F., Ng, J. L., Birima, A. H., Ahmed, A. N. and El-Shafie, A., "Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms," Scientific Reports 12(1), 1-26 (2022).